Busting Myths

Helen Chick
University of Tasmania
<helen.chick@utas.edu.au>

A recent episode of *MythBusters* (Williams, 2013) involved a series of "battles of the sexes" to examine myths and urban legends about things that men are supposedly better (or worse) at doing than women. Some of the processes that were used on the show to investigate these myths, along with the data they generated, can be used to examine some interesting statistical ideas, varying from a quite simple examination of distributions to an elementary exploration of beginning hypothesis testing. As such they would be suitable for secondary classrooms, where the content aligns well with *Australian Curriculum: Mathematics*.

In this article I will describe two of the segments and discuss some of the issues that could be addressed in a classroom. Ideally it would be good to be able to view the segments as part of any lesson (at the time of writing it was possible to purchase the episode via an online media supplier; see note at end of article), but it is hoped that there is enough information in what follows for teachers to be able to provide an explanation of the segments and the data sets should still be sufficient to stimulate good classroom discussion.

Parallel parking

This segment examined the myth that men can parallel park better than women. In order to investigate this issue the presenters got 20 volunteers (10 male and 10 female) to parallel park (i.e., to park a car in the gap between two other cars, nose to tail).

At this point it is useful to stop and think what criteria might be used to identify and actually score a person's parallel parking efforts. This is not the focus of this article but, in the classroom, this issue should be incorporated into the lesson, with a thought-provoking discussion about, "What kind of data can we gather to answer our question, and how can we gather it objectively?"

On the show, the volunteer drivers started off with a score of 100, and lost points for bumping the car in front or behind (scaled by degree of impact), and also lost points for lack of accuracy (too close to either of the other cars, or too far from the kerb). If you have not seen the segment what then followed was an interesting series of lessons in Californian parking conventions: the gap between the original two cars seemed to be quite small by Australian standards, and nudges and outright bumps happened with alarming frequency!

The segment, as presented, actually included some discussion of the partial results and of the developing patterns along the way, hinting at some potentially interesting issues. At the end, the full set of results was shown (see Table 1).

Table 1. Data from the parallel parking experiment, where high values indicate better parking (Williams, 2013).

	1	2	3	4	5	6	7	8	9	10	Average
Men	59	58	34	43	44	31	34	60	33	40	43.6
Women	90	0	68	18	63	72	16	14	61	23	42.5

In a classroom, it might be best to present just the raw data, omitting the pre-calculated means, and then ask the students what conclusions can be drawn and what evidence they have for their claims. Depending on their prior experiences, the students may suggest calculating the means of the men's scores and the women's scores.

Once the means are calculated, the focus question might then be reiterated: "Are men better at parking than women?" Classroom discussion should revolve around how close the two means are to each other, and students could explore what effect changes in one or some of the scores have on the values of the means. For example, you might ask what the mean for the women would be if the score for each of the women went up by just one. Some students may need to calculate this in full; other students will realise immediately that the mean must also increase by one. After some guided exploration and discussion it should be possible to draw the conclusion that the difference between the means is too small to be one that warrants concluding that men are better than women at parallel parking.

Observant students may, however, notice that men's and women's sets of scores are different. The first two scores for the women provide a strong hint about the rest of the values: women score either well or very badly, reflecting a tendency to have either a very limited sense of coordination or a slow and careful approach to parking. Men, on the other hand, have scores that are all relatively close to each other, and not especially high nor yet as low as some of the women, reflecting a tendency to be quick and tolerably accurate (if you allow a minor nudge!).

If we graph the sets of values to show their distribution the striking contrast becomes apparent (see Figure 1). The men's values are all clustered around their mean, and range between 31 and 60. The women's values, on the other hand, show a clearly bimodal distribution, with a cluster of low scores and a cluster of high scores. As it happens, in fact, no score from a woman lies within the range of men's values.

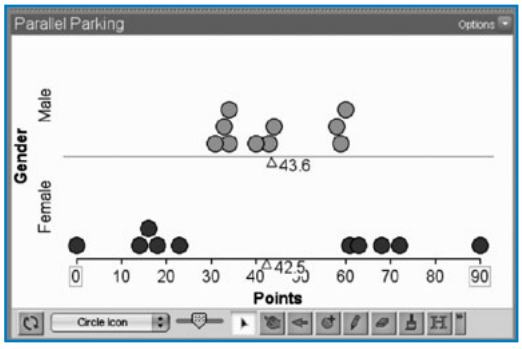


Figure 1. The distributions of men's and women's scores, plotted in *TinkerPlots* (Konold & Miller, 2011).

So, although it may be true to say that on average men and women demonstrate the same parallel parking scores, the difference in the distributions suggest that perhaps the two genders do park differently from each other. The mean certainly does not tell the whole story in this case.

Multitasking

Another segment on the same episode examined a second myth: the claim that women are better at multitasking than men. Again, it would be useful to have classroom discussion about how to design an investigation of this issue. The *MythBusters*' approach was to require participants to complete a collection of tasks such as ironing, getting dressed, preparing lunch, dealing with incoming phone calls, and keeping track of a baby that was not allowed to stray into certain areas, over a 20 minute period. Again, participants started with a score of 100

points and lost points for failing to achieve certain things within the time frame, getting questions wrong over the phone, or losing track of the baby. Ten females and ten males took part in the experiment, and their results are shown in Table 2.

Table 2. Data from the multitasking experiment, where higher values indicate better performance (Williams, 2013).

Multitasking											
Women	100	70	60	80	80	80	80	60	50	60	72
Men	50	80	50	80	90	80	50	60	50	50	64

This time, if you supply only the scores and allow students to decide which group performed better overall, a determination of the means appears to indicate a more clear-cut difference between the genders. Indeed, it seems obvious that women are better at multitasking than men.

But are they? Is that difference of 8 really big enough? Could it have happened by accident rather than as a result of a genuine difference between the groups?

This question of whether a difference is real or big enough is an important one in statistics. The analysis below could be developed in a classroom with the assistance of technology. There are a few steps required, and it will require care to ensure that students understand what data and statistics are being examined at each stage. Nevertheless the analysis has the capacity to help students come to understand whether or not a result is a rare and thus possibly genuine difference, or a common might-have-happened-by-chance difference.

One way of investigating whether or not the result could have happened by accident is to imagine shuffling all the scores, and randomly allocating them to two groups, one labelled men and one labelled women. Our reason for doing this would be to see if a difference between the means that is as big as 8 is likely to happen. If it does not happen very often, then that suggests that there is possibly something special about the fact that it did occur in our actual experiment, and that, as a result, there probably really is a difference between men and women. On the other hand, if a difference of 8 or more occurs fairly regularly, even with a muddled up collection of men and women, this suggests that there was nothing special about our experiment when the men and women are separate because differences of 8 or more are not rare.

So, how do we make up our muddled up set of two groups? Initially this could be done manually: write the original 20 scores on 20 post-it notes, and then toss a coin for each one to decide whether that score is now going to be a 'male' score or a 'female' score. Once there are 10 scores in one of the groups, the remaining unallocated scores go into the other group. We now have ten scores for men and ten scores for women, and we could work out the means of each of the two groups and see how far apart they are. I actually tried this, and my newly allocated group of 'women' had an average score of 71 and the new men had a score of 65, giving a difference between the two groups of +6.

The trouble is, one test is insufficient to tell me whether or not a difference of 8 is actually rare. I could shuffle and allocate my 20 post-it notes again, but I am going to need to do this many times in order to get a more reliable understanding of how common the difference of 8 or more is. Clearly, shuffling post-it notes is tedious; surely technology could come to the rescue. In what follows, the statistical education software package *TinkerPlots* (Konold & Miller, 2011) is used to explore the problem; as an alternative, the Appendix contains a description of how to use *Excel* to reproduce a similar exploration, but in a less self-contained way.

Software like *TinkerPlots* offers the capacity to shuffle data for us, using a Sampler, with a Counter and a Mixer, that allows us to allocate our original 20 scores randomly to males and females. Figure 2 shows the list of already allocated

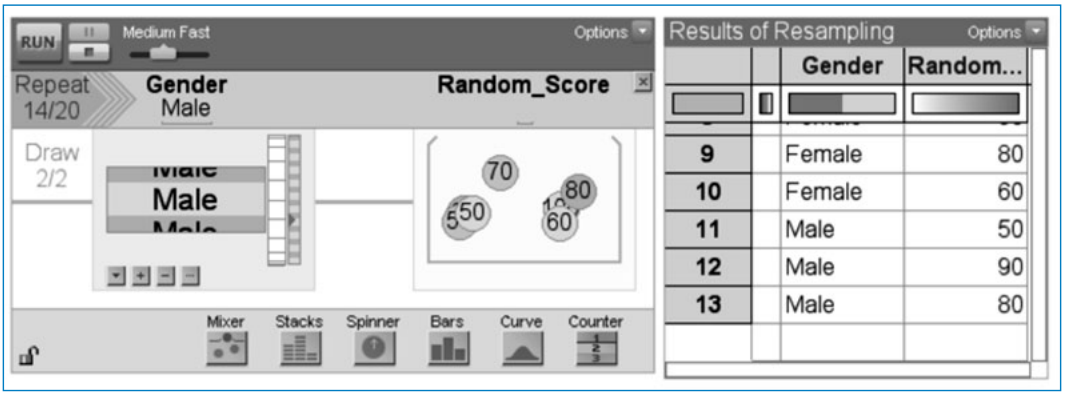
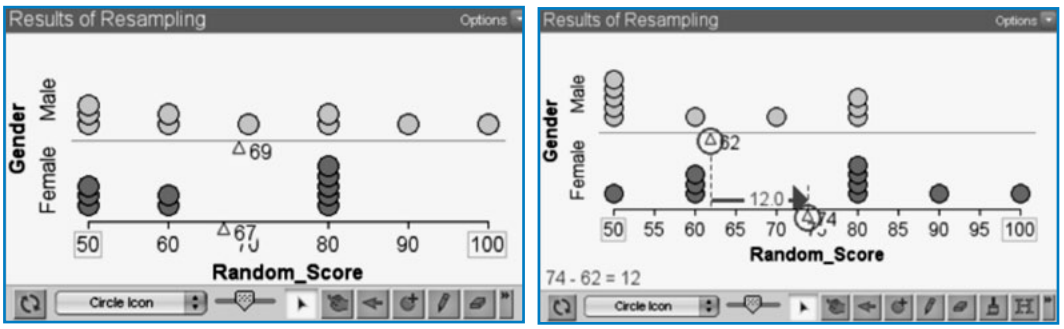


Figure 2. A resampling mixer in TinkerPlots to allow the creation of two groups with randomly allocated scores. On the left, a few scores remain to be assigned to the remaining 'males'. The already assigned scores are shown in the window on the right.

scores on the right, and the remaining scores and to whom they will be assigned on the left. (A complete description of how to set up *TinkerPlots* in order to conduct a resampling activity like this is given in Watson (2013).)

Once the scores have been allocated, *TinkerPlots* can be used to graph the 'male' and 'female' results, as well as display the means. Figures 3a and 3b show the results from two resamplings, each obtained by shuffling our original 20 scores and allocating them randomly to 'males' and 'females'. As can be seen, in Figure 3a the mean for the 'males' actually ends up being 2 higher than the score for the 'females', whereas in Figure 3b the 'females' outperform the 'males' by 12. In Figure 3b we have set up *TinkerPlots* so that it shows and calculates this difference between the means (see the circles and arrows). This is useful, because it is this difference that is of interest to us. In fact, what we really want to do is lots and lots of resampling, and keep track of the differences between the means that occur each time.

In Figure 4, we have now set up *TinkerPlots* so that it keeps a record of the differences between the pairs of means, for a series of shuffles/resampling. The



Figures 3a and 3b. Means and distributions of scores from two resamplings that randomly allocated the original 20 scores to 'male' and 'female'.

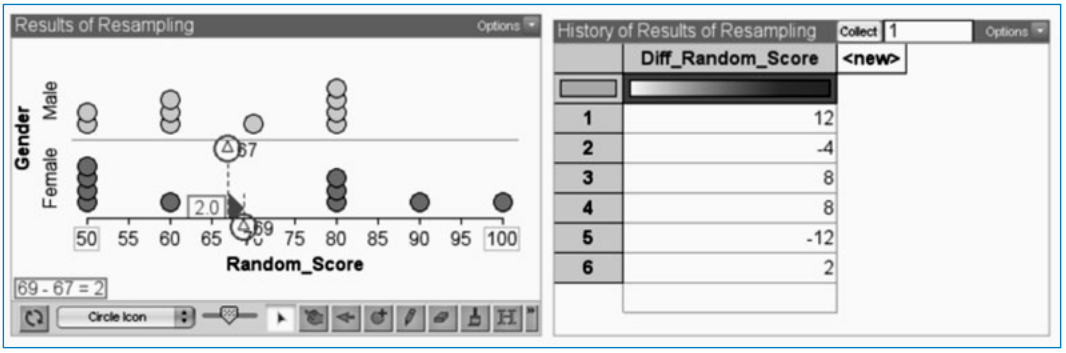


Figure 4. The results of another resampling on the left, and, on the right, a list of the differences between the means for a succession of resampling shuffles.

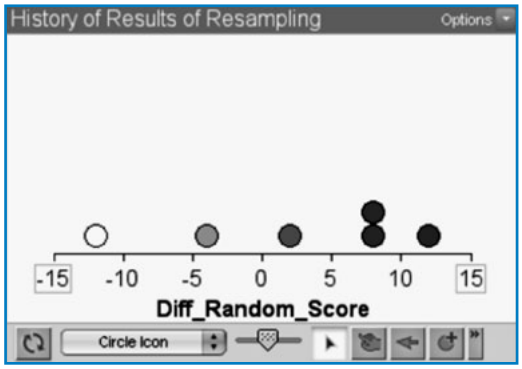


Figure 5. Stacked dot plot, showing the differences between the means for six sets of resampled shuffles (this shows the data from the right of Figure 6).

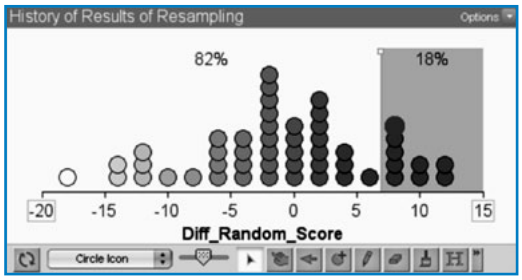


Figure 6. The distribution of differences in the means for 50 random allocations of our original 20 scores.

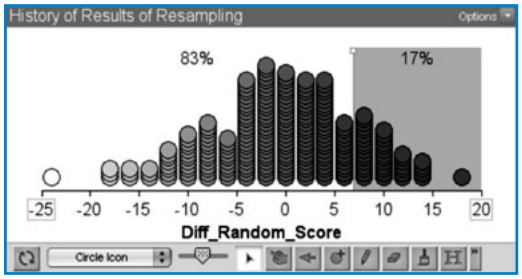


Figure 7. The distribution of differences in the means for 250 shuffles of our original 20 scores. (This was actually a separate run of 250 rather than a continuation of the 50 in Figure 8.)

results from six shuffles are shown on the right, with the actual distribution and means of the sixth shuffling shown on the left. As you can see from the list of differences, sometimes the 'females' beat the 'males' and sometimes it is the other way around (when the differences are negative). In these few shuffles, a difference of 8 between the 'females' and the 'males' actually happens a couple of times, but we still do not have enough data to have a sense of whether it really is a common or rare event from our randomly allocated scores.

What we can start to do, however, is to graph these results. Figure 5 shows a stacked dot plot of the set of six differences that we have already found from our six resampling shuffles. What we need to do next is to let *TinkerPlots* complete a much larger number of resampling shuffles so we can see what happens to this graph, and answer the question about whether or not a difference of 8 is rare.

In Figure 6, *TinkerPlots* has resampled 50 times; in Figure 7, there are 250 resamples. We could let it do even more to confirm the trend that is becoming apparent. *TinkerPlots* allows you to highlight sections of the graph, and can tell you how many data values are in the highlighted region. In each of Figures 6 and 7 we have highlighted the differences that were 8 or above, i.e., where 'females' outscored 'males' by 8 or more. As you can see, around 17% of the time a random allocation of our original set of scores will result in groups of 'males' and 'females' that have means that differ by 8 in favour of the 'females'.

The question is, is 17% of the time 'rare'? A likelihood of 17% means that having a difference of 8 or more in favour of women is about as likely as rolling a 2 on a die, which we know happens $\frac{1}{6}$ of the time. Although some students may think that getting a 2 is uncommon—and there are some important underlying probability understandings that need to be addressed about this issue, which are not the scope of this article—it is not really a rare event. It happens 'every so often'; so we get a 2 from a die about every 6 rolls. This means that for our randomly allocated groups we get a difference of 8 or more about every six shuffles. The point is that this difference occurs relatively frequently even when our two groups have been constructed totally randomly. Thus the outcome of a difference of 8 between the means of our original groups is not a particularly unusual thing, and so such a difference could be just happening by chance and not because the groups were gender based.

So, with a difference of 8 or more turning out to be relatively common, we have to conclude that the difference between the performance scores of the men and the women could be due merely to chance. It is not really rare enough to conclude that it is a genuine difference between men and women. The *MythBusters* program, which only examined the two means and not the relative rarity of the difference between them, claimed that women are better than men at multitasking, but our

analysis says that the difference could just as easily be due to chance and that there is not enough evidence to say women really are better than men.

Conclusion

These two sets of data provide a good opportunity to explore how the careful use of statistics can help us to understand a situation: to identify differences and to be cautious about how significant or 'real' those differences might be. The first example shows how the simple yet powerful technique of examining a distribution can give us a better picture of a situation than solely using measures of central tendency such as the mean, valuable though these are. Secondly, although we have not formally entered the realm of significance testing, the analysis done for the multitasking example gives us a better sense of whether or not a 'found result' is actually rare—and so likely to show a genuine difference between groups—or sufficiently common that it could well have occurred just by chance. Examples such as those found in this particular *MythBusters* episode provide a motivating opportunity to examine such statistical issues.

Acknowledgements

The author thanks Jane Watson for useful discussions and assistance in producing the *TinkerPlots* graphs.

Note

At the time of writing, this episode of *MythBusters* could be purchased from an online media store such as iTunes for about $3.50. It is contained within Season 8 Volume 1, where it is episode 11. It can be purchased as an individual episode.

References

Konold, C., & Miller, C. D. (2011). *TinkerPlots: Dynamic data exploration* [computer software, version 2.0]. Emeryville, CA: Key Curriculum Press.

Watson, J. (2013). Resampling with TinkerPlots. *Teaching Statistics, 35*(1), 32–36.

Williams, C. (Writer). (2013). Battle of the sexes – Round 2 [Television series episode]. In J. Hyneman, A. Savage & D. Tapster (Executive producers), *MythBusters*. San Francisco, CA; Beyond Entertainment Ltd.

Appendix: Using Excel to explore the reshuffling of the multitasking data

There are two things that we need to get *Excel* to do, in order to explore the same things that we were able to do with the resampling process in *TinkerPlots*: (i) we need to get it to reshuffle the data so that we get two groups where we can work out the means, and (ii) we need to accumulate the collection of values of the difference between the means so we can determine if the difference of 8 is rare or common. As it happens, (i) is easy to do, but (ii) has to be done manually.

First, enter the data into *Excel*; in Figure 10 I have put the scores in column B and the gender in column C, with the men first and the women second. By recording the genders in column C it is possible to resort the data quickly back into the separated groups of men and women, even after you have done some reshuffling. I have used Cells B23 and B24 to calculate the means of the first ten and second ten values by entering "=AVERAGE(B1:B10)" in B23 and "=AVERAGE(B11:B20)" in cell B24. With the data sorted by gender this will show the original means of the men (64) and women (72).

	A	B	C	D
1	0.35774227	80	M	
2	0.74306992	60	M	
3	0.97653981	80	M	
4	0.69216247	50	M	
5	0.15936582	80	M	
16	0.05389385	50	W	
17	0.86537766	70	W	
18	0.56304586	100	W	
19	0.63379968	80	W	
20	0.73814965	60	W	
21				
22				
23		64	First 10 average	
24		72	Second 10 average	

Figure 8. The original data, in order, in an Excel spreadsheet.

	A	B	C	D
1	0.59010564	80	M	
2	0.61900557	100	W	
3	0.14672985	80	W	
4	0.74833811	70	W	
5	0.50150714	90	M	
16	0.2714719	80	M	
17	0.31556207	60	M	
18	0.85691706	50	M	
19	0.05352036	80	W	
20	0.40186583	50	M	
21				
22				
23		74	First 10 average	
24		62	Second 10 average	

Figure 11. Shuffled data values.

The next step is to find a way to shuffle the data, so that the values can be randomly allocated to two groups. To do this, use the random number function to put a random number in each of the first 20 cells of column A, so that each data value also has a random number associated with it (see Figure 8). To do this, use "=RAND()" in each cell, which produces a randomly generated number between 0 and 1.

Having done this, select the first 20 rows of the spreadsheet (in particular, do not select the 23rd and 24th rows with the average calculations in them). Now use *Excel*'s Sort facility to sort the 20 data values based on Column A. This will use the random numbers to sort the data (see Figure 9, noting that in the process of performing the Sort, *Excel* will also re-evaluate the random numbers in Column A, and so it does not actually look as if the Sort has taken place, except for the fact that the men's and women's values in Columns B and C are now muddled up).

You can now regard the first ten values to be our men and the second ten values to be our women. (It should be noted that there is small potential for confusion here. Because I wanted to keep the original data accessible, each score still has the gender of its original owner attached to it. These values in Column C have to be ignored; what we are doing here is designating that the first 10 of our shuffled people are men—regardless of what they were originally—and the last 10 are designated as women. We will then work out the average scores for each of these newly determined groups.) The averages will be calculated automatically in cells B23 and B24, because their calculations are based on the first 10 in the list, who are currently assigned to be men, and the remaining 10, who are currently assigned to be women. By repeatedly Sorting—since the random values are re-generated each time without having to do anything extra—new shuffles can be generated, and each time the resulting new means of the men and women will be shown.

Unfortunately, keeping track of these differences needs to be done manually. Figure 9 shows that the difference between the 'men' (the first 10 values, with an average 74) and the 'women' (average 62) is –12, because in this case it is the 'men' ahead of the 'women'. By repeatedly shuffling and keeping a tally of how many shuffles have been conducted and how many times that 'women' have a mean value that is 8 or more larger than the mean value for the 'men' it is possible to show that this event occurs about 17% of the time.